July 30, 2025

Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York

Re: *In re: OpenAI, Inc., Copyright Infringement Litigation*, 1:25-md-03143 (SHS) (OTW); this document relates to No. 1:23-cv-11195 (SHS) (OTW)

Dear Magistrate Judge Wang:

I write on behalf of News Plaintiffs[1] to request entry of News Plaintiffs' proposed order, attached hereto as Exhibit A[2], regarding a sampling and search protocol for the API and ChatGPT output log data OpenAI has recently been ordered to preserve.[3] *See MDL* Dkt. 33, p. 2 (ordering OpenAI to "preserve and segregate all output log data that would otherwise be deleted on a going forward basis"); *MDL* Dkt. 79. While the parties have reached agreement on the search terms and certain sample populations for the ChatGPT data, three outstanding disputes remain: (1) OpenAI's refusal to include the recently preserved API output log data in the sampling and search protocol, which News Plaintiffs believe is a necessary part of conducting a Rule 37(e) analysis of OpenAI's prior destruction of such data; (2) OpenAI's restrictions on the use of the hit reports and related analyses in the litigation; and (3) OpenAI's request to recover certain costs associated with its retention of the consumer output log data. Prompt resolution of these issues will enable the parties to expeditiously complete the search and sampling analysis directed by this Court and engage in further discussions regarding the Rule 37(e) analysis with respect to OpenAI's deletion of over 8 billion consumer ChatGPT conversations and virtually all of its historical API conversation data.[4]

I. Background

For several weeks, News Plaintiffs have been working to reach agreement with OpenAI on the best way to sample the output log data subject to this Court's preservation orders and conclude this analysis expeditiously. This has been productive, as the parties have agreed that the analysis should include three sample populations of ChatGPT "conversation" data. *See* Exhibit A, paragraph 1. Each sample population will include five million rows in the tables of output log data OpenAI has been preserving. *Id*.

Following the June 26, 2025 settlement conference, the parties also reached resolution on several outstanding issues: (1) search terms, including News Plaintiffs' narrowed "news" search term and list of domains containing AI-generated journalism, (2) exclusion of certain non-US data

[1] This motion is filed on behalf of The New York Times Company ("The Times"), the *Daily News* Plaintiffs, and the Center for Investigative Reporting.
[2] Exhibit B includes a redline of the proposed order as compared to OpenAI's July 25 proposal.
[3] News Plaintiffs have separately filed a motion regarding the sampling size to be used for historical ChatGPT output logs (Dkt. 394), but this motion addresses the sampling analysis necessary to address OpenAI's failure to preserve certain output logs.
[4] News Plaintiffs proposed coordinating a joint filing on these issues to avoid burdening this Court with cross-motion practice, but OpenAI declined.

in view of OpenAI's stated privacy concerns, and (3) OpenAI's production of classifier information for historical API data.

In view of the parties' agreement thus far, News Plaintiffs believe that OpenAI is able to begin sampling and running the search terms over the agreed-upon ChatGPT sample populations. However, in order to complete this process, the parties require the Court's resolution of the following three disputes.

II. API Data

OpenAI has refused to sample and search the preserved API data, and has represented that it does not have classifier data for the API data that it has been ordered to preserve. Accordingly, News Plaintiffs submit that a fourth sample population should be included, consisting solely of API data and also including five million rows from OpenAI's tables, for at least two reasons.

First, News Plaintiffs have been requesting a sample of the API data for over a year as part of their merits-based analysis for the case, which is relevant to *inter alia*, OpenAI's infringement of News Plaintiffs' copyrights due to "regurgitation," dilution of News Plaintiffs' trademarks due to "hallucinations," and market dilution of News Plaintiffs' copyrighted works due to the proliferation of AI-generated news content. There is also reason to believe that the API data will differ significantly from ChatGPT output, including because – as explained during the Technology Tutorial – the API permits users to reduce the "temperature" so that its outputs are more likely to generate verbatim memorized content. *See* Ex. C at 202:17-203:6. Both The Times and the *Daily News* plaintiffs served requests for inspection and document production related to OpenAI's API output logs in May and June 2024. *See Times* Dkts. 379-5, 379-6, 379-7 and 379-8. The parties met and conferred numerous times about these requests, and OpenAI represented it was investigating its ability to search its API output logs. On March 19, 2025, OpenAI informed News Plaintiffs that "[o]n March 5, 2025, there were over 100 billion individual API completions stored" (Dkt. 43-18), and on March 31, 2025, OpenAI produced a limited sample of its API output logs to facilitate discussions. But after nearly 11 months of discovery on this issue, it was not until an April 9, 2025 call with OpenAI's technical consultants regarding the output logs that OpenAI first informed News Plaintiffs they would not make ***any*** API data available. Dkt. 43-8. News Plaintiffs pushed back, and made clear that they continued to seek the API output log data they had been discussing with OpenAI since the summer of 2024. *Id.* In view of News Plaintiffs learning of OpenAI's deletion of virtually all historical API data, News Plaintiffs have since requested that OpenAI produce a sample of API data from the API data subject to this Court's preservation order.

Second, running a search over the API data is necessary to inform the Rule 37(e) analysis. News Plaintiffs requested that the API data be included as part of the search and sampling exercise in a filing before the June 26 settlement conference, raised the issue during the settlement conference, and continued to ask that the API data be included following that conference. *See MDL* Dkt. 263 at 2-3; Ex. D. News Plaintiffs are prejudiced by the unavailability of virtually all of the historical API output log data, which, as explained above, would have contained direct evidence of copyright infringement and trademark dilution and rebutted OpenAI's assertions about

the way their models operate. While News Plaintiffs explored using classifier data to assess this issue, OpenAI has represented that it does not have classifier data for the API data it has been ordered to retain. Accordingly, the most practical and expedient way to analyze the relevance of the preserved API data is to sample and search it. An order compelling OpenAI to run the agreed-upon search terms over the API data will provide information that will help News Plaintiffs and this Court assess the relevance of the API data to the Rule 37(e) analysis, including, for example, the current prevalence of prompts requesting news content and outputs providing such content. It may also be informative to see the prevalence of prompts and outputs referencing the News Plaintiffs, although there may be a highly reduced universe of such instances because, for more than a year, OpenAI has instituted News Plaintiff-specific blocks.

In view of the extensive discussions, and likely further delay in getting any such sample of API data, OpenAI should be ordered to promptly run the keyword search over the API data by no later than August 15.

III. Use of the Sampling and Search Exercise in the Litigation

OpenAI seeks to limit the use of the sampling exercise solely to the Rule 37(e) analysis at this time. Such a limit is inappropriate because – as explained above – the API log data is relevant to the merits of this case. Under OpenAI's theory, if the sampling process reveals a high prevalence of news-related output or output containing or referencing News Plaintiffs' intellectual property, for example, the News Plaintiffs would be prohibited from showing that evidence to their experts or introducing it at summary judgment or trial. OpenAI has provided no legal basis to justify this limitation, and there is none. It is also premature to propose any limits on how the results may be used before the parties have analyzed the results, and such results may be important to show a jury what data was lost. For example, appropriate measures to cure prejudice may include "permitting the parties to present evidence and argument to the jury regarding the loss of information, or giving the jury instructions to assist in its evaluation of such evidence or argument." 2015 Advisory Notes to Rule 37(e).

IV. Cost Shifting

OpenAI has insisted on a cost-shifting provision in its version of the proposed order. This is inappropriate. The parties are conducting this sampling exercise because OpenAI deleted significant portions of its output logs, so if any party is subject to an undue burden by this process, it is the News Plaintiffs. Nothing in the record justifies shifting the usual presumption that a party must bear its own discovery costs, particularly in these circumstances. *See* MDL Dkt. 262 at 6; *see Treppel v. Biovail Corp.*, 233 F.R.D. 363, 372-373 (S.D.N.Y. 2006) ("the presumption is that the party possessing information must bear the expense of preserving it for litigation").

***

News Plaintiffs will be prepared to address these issues at the August 12 Hearing.

July 30, 2025

Respectfully submitted,

/s/ Steven Lieberman
Steven Lieberman
Rothwell, Figg, Ernst & Manbeck, P.C.

/s/ Ian Crosby
Ian B. Crosby
Susman Godfrey L.L.P.

/s/ Stephen Stich Match
Stephen Stich Match
Loevy & Loevy

cc: All Counsel of Record (via ECF)